PLISKA

STUDIA MATHEMATICA BULGARICA

ПЛИСКА

БЪЛГАРСКИ МАТЕМАТИЧЕСКИ СТУДИИ

Pliska Stud. Math. Bulgar. **16** (2004), 291-298

PLISKA
STUDIA MATHEMATICA
BULGARICA

INTERACTIVE STEPWISE DISCRIMINANT ANALYSIS IN MATLAB

D. L. Vandev[1]

The program `ldagui.m` is an interactive tool for linear and quadratic discriminant analysis. The reason for developing such a tool consists in failing of conformity with conventional statistical programs in following aspects: treating of missing data; interaction with the user; testing the quality of obtained models.

1. Overview

Discriminant analysis (DA) is a very popular tool in applied statistics. The program `ldagui.m` is developed in the frame of MATLAB. It is used with the help of menus, shortcuts, listboxes and a slider.

- First shortly the mathematics behind DA will be outlined.
- Then the menus and shortcuts of the program will be described.

2. Discriminant Analysis

Let us suppose that two random variables were observed:

1. *continuous* ξ with values $x \in R^p$;

[1]Partially supported by project PRO-ENBIS: GTC1-2001-43031 and WINE DB: G6RD-CT-2001-00646.

2000 *Mathematics Subject Classification:* 62-04, 62H30, 62J20

Key words: stepwise discriminant analysis linear quadratic MATLAB

2. *discrete* (or categorical) η with values $y \in \{1, 2, \ldots, G\}$.

They have joined distribution (DA model):

- $\mathbf{P}(\eta = y) = p_y$
- The conditional distribution of $\xi \in R^p$ given by $\eta = y$ is described by the density of Gauss distribution $p_y(x) = \varphi(x, \mu_y, C_y)$ described by two parameters: mean – μ_y and covariance – C_y.

Suppose that the parameters of this model are known. That is: the *prior* probabilities – $\{p_y\}$; *group means* – $\{\mu_y\}$; within group *covariance matrices* – C_y.

Then according to the famous formula of Bayes the conditional probability of $\eta = y$ given by $\xi = x$ can be written down:

$$\mathbf{P}(\eta = y|\xi = y) = q(y|x) \stackrel{def}{=} c(x)p_y(x) = c(x)\varphi(x, \mu_y, C_y), \tag{1}$$

where $c(x)$ is a normalizing constant, such that $\sum_y q(y|x) = 1$. This probability is called *posterior* and it means that the observation x belongs to the group y with probability $q(y|x)$. Then according to the maximum likelihood principle the classification rule becomes:

$$\widehat{y}(x) = \underset{h}{argmax} : q(h|x). \tag{2}$$

2.1. Linear and Quadratic DA

Suppose that within group covariances C_y are equal:

$$C_y = C, \quad (y = 1, 2, \ldots, G). \tag{3}$$

Then the maximum likelihood rule (2) becomes a set of inequalities:

$$p_{\widehat{y}} f(x, \mu_{\widehat{y}}, C) \geq p_h f(x, \mu_h, C), \quad (h = 1, 2, \ldots, G) \tag{4}$$

or (what is the same):

$$L_{\widehat{y}}(x) = b(\widehat{y})'x + a(\widehat{y}) \geq L_h(x) = b(h)'x + a(h). \quad (h = 1, 2, \ldots, G) \tag{5}$$

The observation x should belong to the group y, if for each h the inequality (5) holds. The linear functions L are called *discriminant functions.*

If the assumption (3): $C_y = C$ is not appropriate, the corresponding functions become quadratic.

The estimators of common $\widehat{C}$ and within group $\widehat{C}_y$ covariance matrices are constructed in the usual way. In `ldagui.m` the estimators $\widehat{C}_y$ will be stabilized using optimal value of α :

$$\widehat{\widehat{C}}_y = (1-\alpha)\widehat{C} + \alpha\widehat{C}_y. \tag{6}$$

As criteria for this optimization the classification or the theoretical error of the DA model can be used. This variant of DA is called QLDA.

2.2. Theoretical error

By definition theoretical error of a DA model is the probability of incorrect classification of a random observation. In `ldagui.m` a huge sample with 6000 observations per group is produced according estimated within group means and covariance matrices. The sample is classified and results are reported on the MATLAB command window. This may be considered as an estimate of the theoretical error of a model with equal prior probabilities.

2.3. Selecting variables

The standard approach of Fisher is to maximize the between group variance or (what is the same) to minimize the common within group variance:

$$SS = \sum_{g\in G}\sum_{i\in I(g)} (x_i - m(g))(x_i - m(g))'.$$

Trace or determinant can be used to find corresponding variables. Now in all programs the so called Wilk's lambda is used:

$$\Lambda = \frac{\det(SS_{in})}{\det(SS_{total})}.$$

It is easy to calculate (see [Jennrich, 1977]) and convenient to update if a variable is to be entered into (or to be removed from) the model.

3. Menus

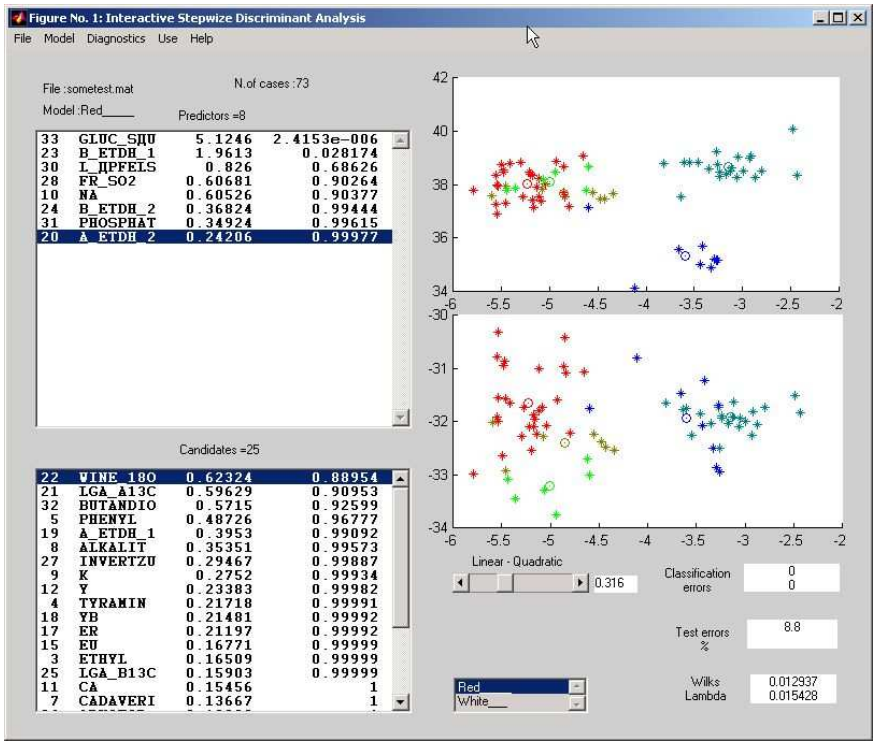


Figure 1: The main window of the program

Before the menus of the program will be considered we need some explanations of the words used.

Comma separated values

These (.csv) files are common for many applications. They are easily exported and imported by Excel and Statistica programs.

`ldagui.m` assumes that:

- the separating character is comma;
- the first row contains strings for variable names;
- the first column contains strings for case names.

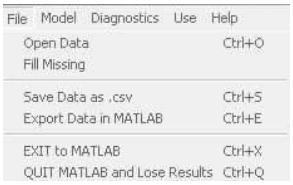
No commas should be in these strings. All the other fields should contain numbers (or should be empty for missing values).

Categorical variables

The categorical variables should have consecutive positive integer values. If exporting them from Statistica you should say integers instead of text values.

Any information about text values which were created in Statistica will be lost and moreover their values in `ldagui.m` may be changed to first natural numbers: 1,2,3...

3.1. File



The File drop down menu may be used separately in order to fill the missing data with within group means.

Figure 2: File menu

Open Data	The program loads a data (.csv) file. The program will ask you to select one obligatory variable for the classification. The use of `ldagui.m` is impossible without classification variable.
Filling Missing Data	You will be asked to choose a (not obligatory) selection variable. The classification and the selection variable will be used in the algorithm for filling missing data. Filling Missing Data is done automatically by `ldagui.m upon reading of .csv data. They are replaced by within group means. These means are formed by each combination of values of classification and selection variables.`
Save Data	Saves the data file in a form of a comma separated file for the later import into Excel or Statistica.
Exit to MATLAB	Saves the MATLAB workspace in a tempmodel.mat for the following examination and use.
Quit MATLAB	Simply closes MATLAB

3.2. Model

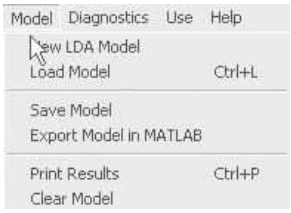


Figure 3: Model menu

Under model here will be understood:

1. The training sample (with no missing values);
2. A subset of cases having fixed value of the selection variable (if any exist);
3. A subset of variables chosen for predictors (may be empty);
4. The fixed value of the parameter α of nonlinearity;
5. The estimated parameters of the DA model.

Build Model — Performs all preliminary calculations for an empty model with no selection variable taken into account. To activate this option click on the Selection Listbox.

Load Model — Loads previously saved model (workspace).

Save Model — Saves the current model with data, names, selected groups, predictors, etc. for the later use. In fact the current workspace of MATLAB is saved.

Print Results — The following results are printed in the MATLAB command window:

- File name – the name of the file (data or model) you have loaded recently;
- Model name – the name of the corresponding value of the selection variable;
- Number of cases in the training sample.
- Variables in model with their $F-$ and $p-$ values ordered;
- Value of the parameter α responsible for nonlinearity;
- Value and p-value of Wilk's Λ;
- Results of the classification of the training sample – number of errors;
- Cases classified with probability below .8;
- Estimated error of the model by groups.

Clear Model — Clears any information for the model. You should start with Build model step.

3.3. Diagnostics

Figure 4: Diagnostics menu

The tools proposed for making adequate decision are:

- Test – (Ctrl-t) – produces a test random sample;
- Leave-One-Out – checks the model against deleting each of observations;
- Plot – (Ctrl-g) – makes two plots over canonical variables.

Test
: A small sample with 100 observations per group is produced according estimated within group means and covariance matrices. The sample is classified and the results are reported on the MATLAB command window. This may be considered as an estimate of the error of the model. This step can be repeated to be sure or the print menu with larger samples can be used 3.2..

Leave-One-Out
: The following standard procedure is performed:

1. For each observation in the training sample a model with the same variables will be build but without this particular observation.

2. The training sample will be classified with this new model and classification errors are counted.

3. The errors for all observations will be summarized and will be reported.

Plot
: Second (upper plot) and third canonical variables will be plotted against the first (on horizontal axes). The training sample will be described by different colours for the groups.

3.4. Use

This menu aims to help the user to apply the model to some other data set. Below this set is called sample.

Figure 5: Use menu

Load sample	A standard data (.csv) file is loaded which should not contain missing values in the columns used for recognition. Columns to use should have the same variable names.
Print results	Results of classification are printed.
Save sample	The sample is saved in a data (.csv) file with resulting classification in the first column.

4. Algorithms

The calculations are based on the paper of [Jennrich, 1977] in the classical collection of [Einslein et al., 1977] being in the foundations of the package BMDP (see [Dixon, 1981]).

REFERENCES

[Dixon, 1981] J. Dixon (ed.) *BMDP Statistical Software – 81*, 1981, University of California, Los Angeles.

[Vandev, Römisch, 2003] D. Vandev and U. Römisch. Comparing several methods of Discriminant Analysis on the case of Wine Data. *Pliska Stud. Math. Bulg.* (Eds D. Vandev, N. Yanev) Proceedings of the Seminar on Statistical Data Analysis 2003. Sozopol, 21–28.06.2003, **16**, 299–308.

[Einslein et al., 1977] K. Einslein, A. Ralston, et al. (eds) *Statistical Methods for Digital Computers*, 1977, John Wiley & Sons, New York.

[Jennrich, 1977] R. I. Jennrich. Stepwise discriminant analysis. In: *Statistical Methods for Digital Computers*, 1977, John Wiley & Sons, New York, 76–95.

D. L. Vandev
Sofia University "St.Kliment Ohridski"
Sofia, Bulgaria
e-mail: vandev@fmi.uni-sofia.bg